DOCUMENT RESUME

ED 334 235 TM 016 759

AUTHOR De Champlain, Andre; Gessaroli, Marc E.
TITLE Assessing Test Dimensionality Using an Index Based on Nonlinear Factor Analysis.
PUB DATE Apr 91
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Equations (Mathematics); *Factor Analysis; Foreign Countries; *Item Response Theory; Mathematical Models; *Sample Size; Simulation; Test Format; *Test Items
IDENTIFIERS *Incremental Fit Index; *Nonlinear Models; Stouts Procedure; Unidimensionality (Tests)

ABSTRACT

A new index for assessing the dimensionality underlying a set of test items was investigated. The incremental fit index (IFI) is based on the sum of squares of the residual covariances. Purposes of the study were to: (1) examine the distribution of the IFI in the null situation, with truly unidimensional data; (2) examine the rejection rate of the IFI under various simulation conditions of a two-dimensional test structure; and (3) compare the performance of the IFI with the T-statistic of W. Stout (1987). Data sets were computer-generated for sample sizes of 500 and 1,000 with test lengths of 15 and 45 items each. The IFI based on the sum of the squares of the residual covariances of the one-dimensional and two-dimensional non-linear factor analyses of dichotomous test data did show fairly high rejection rates of unidimensionality when two-dimensional data were generated. The results suggest that the statistic has the potential for use in the assessment of unidimensionality of test data and in the determination of the number of dimensions underlying a test. The T-statistic seemed best suited for long tests having large sample sizes, while the IFI might be preferable for smaller test lengths or smaller samples. Five tables present study data. A 23-item list of references is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED334235

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
ANDRE DE CHAMPLAIN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Assessing Test Dimensionality Using An Index Based On Nonlinear Factor Analysis

Andre De Champlain
Marc E. Gessaroli[1]

Faculty of Education
University of Ottawa

Presented at the annual meeting of the American Educational Research Association, April 4, 1991, Chicago, IL

Running Head: IFI Test of Dimensionality

1. The authors would like to thank J. Boulet and B. Zumbo for their help and suggestions.

BEST COPY AVAILABLE

TM016759

Assessing Test Dimensionality Using An Index Based on Nonlinear Factor Analysis

The numerous studies dealing with Item Response Theory (IRT) that have dominated the measurement literature in the past decade attest to its importance in the development and analysis of tests and items. Its many advantages, namely that it is sample free and provides the test developer with information pertaining to a wide range of examinee abilities, have generated considerable interest in the area of educational testing. However, its widespread application has been hindered by strong assumptions underlying IRT models, especially the requirement that the underlying trait be unidimensional. This assumption, however, is often unreasonable in practical testing situations. A mathematics test, for example, entails not only mathematical ability but also the ability to read and understand the problems being presented. In addition, authors that have estimated the robustness of IRT item and ability parameter estimates obtained from multidimensional data generally have shown that these values are unreliable, most notably, when several equally important abilities are required to correctly answer an item (Ackerman, 1987; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Reckase, 1979, 1986).

This important consideration has lead to the development of

statistical techniques to assess test dimensionality or, more realistically, departure from the assumption of unidimensionality. The majority of the research in this field has focused primarily on the evaluation and/or development of indices based on principal components analysis(PCA)/common linear factor analysis(LFA), the Holland-Rosenbaum procedure, Stout's essential dimensionality and residual covariance analyses.

The first group of studies typically has examined the extent to which those indices derived from PCA/LFA based on phi and tetrachoric correlation matrices (e.g., % of variance explained by the first component, scree plots, ratio of first to second eigenvalue, etc,.) could be helpful when assessing the dimensionality of dichotomous data generated from a logistic model. The results obtained in these studies diverged greatly depending on the characteristics of examinees/items and were generally quite unreliable in identifying the correct number of dimensions underlying a simulated data set. Generally, these indices tend to overestimate the number of components/factors underlying the items (Berger & Knol, 1990; De Ayala & Hertzog, 1989; Hambleton & Rovinelli, 1986; Hattie, 1984; Zwick & Velicer, 1986). In addition, factor analysis of phi matrices may lead to spurious factors (Green, 1983; McDonald & Ahlawat, 1974). In summary, research in this area appears to discourage the use of indices based on PCA or common LFA. These results are not surprising given the misfit that is to be expected when trying to

fit a linear model to data which conform to a nonlinear (logistic) model.

Rosenbaum (1984) states that if ICCs are monotone nondecreasing functions of a single ability, the local independence of item responses implies nonnegative conditional covariance between all pairs of item responses. Rosenbaum's procedure (1984) therefore tests the assumptions of conditional local independence and monotonicity of item response functions using the Mantel-Haenzel z statistic. Results obtained by Zwick (1987) and Ben-Simon & Cohen (1990) show that the procedure is too conservative. However, the latter authors did obtain encouraging results using a modified version of the procedure that incorporated parallel analysis.

Stout's procedure is based on a new definition of dimensionality: <u>essential dimensionality</u>. He argued that it was unrealistic to believe that a test could truly be unidimensional (i.e., zero residual covariances between items after fitting a one-factor model). Essential dimensionality corresponds to the number of dimensions necessary to satisfy the assumptions of essential independence (i.e., the mean conditional residual covariance which tends towards a minimum as the number of items increases). A test consisting of items U_j, (j=1,...,N) of length N is said to be essentially unidimensional if there exists a latent trait θ such that for all values of θ,

$$\frac{1}{N(N-1)} \sum_{1 \le i \ne j \le N} |Cov(U_i, U_j | \theta)| = 0. \qquad (1)$$

The assumption of essential independence is then tested using the T statistic (Stout, 1987). In addition, Nandakumar (1987) proposed a correction method for the procedure used to calculate the T statistic in order to reduce bias due to easy items being solely included in the first assessment test (AT1). Readers interested in obtaining a more detailed description of Stout's T statistic and the bias correction method should refer to Stout (1987) and Nandakumar (1987). Results indicate that the T statistic appears to be accurate (Stout, 1987), especially when Nandakumar's modification is utilized (Nandakumar, 1987; 1988; 1989). However, the precision of the index seems to decrease as the test length decreases. This procedure should not be used with a small number of items (less than 25) (Nandakumar, 1987).

Another approach quickly gaining popularity is one that treats IRT as a special case of nonlinear factor analysis (see McDonald, 1967, for some of the first work in this area. More recent discussions of this topic can be found in Goldstein & Wood (1989) and McDonald (1989)). Takane & De Leeuw (1987) have shown that the models used in IRT and nonlinear factor analysis are mathematically equivalent, a fact previously alluded to by McDonald (1967). Using this IRT-Factor analysis relationship,

some authors have suggested that the only logical method of assessing dimensionality would have to be based on an analysis of the residual covariance matrix after some type of nonlinear factor analysis (Hambleton & Rovinelli, 1986; Hattie, 1984; McDonald, 1989). Indeed, local independence and unidimensionality of the latent trait would theoretically imply zero residual covariances between all pairs of items at fixed ability levels (i.e., the single ability would account for covariations between items). Results obtained by Hambleton & Rovinelli (1986) as well as Hattie (1984) show that various indices such as the sum of absolute residual covariances and the mean standardized residual correlation tend to be related to the number of dimensions underlying a set of test items. Recently, in keeping in line with Stout's philosophy of essential dimensionality (see equation 1), the mean absolute residual covariance, has been investigated by Berger & Knol (1990) in a simulation study with quite promising results. From a practical perspective, however, the unrealistic test length of the data sets generated (15 items) as well as the small number of replications (10) indicate that the authors' conclusions should be interpreted cautiously and that the index should be assessed in more varied situations before any definite judgment is made about its effectiveness. Also, another possible weakness is that the mean absolute residual covariance is not based on the criterion minimized in the unweighted least-squares estimation

which is used in the nonlinear factor analysis program, NOHARMII (Fraser, 1983).

A new index that is based on the sum of squares of the residual covariances (SSRes) is investigated in this study. The SS_{Res} is the criterion minimized in the ULS estimation procedure used in NOHARMII (Fraser, 1983). Specifically, the index proposed in this study is an incremental fit index (IFI). In the context of assessing the dimensionality of a set of test items, we can define the IFI as:

$$IFI_m = \frac{SS_{Res}(m\text{-}factor) - SS_{Res}((m+1)\text{-}factor)}{SS_{Res}(m\text{-}factor)} \qquad (2)$$

The IFI calculates the proportion of the sum of squares of the residual covariances from the m-factor solution that is accounted for by the (m+1)-factor. If the (m+1)-factor) is important in explaining the structure of the items, then the IFI should be quite large.

The theoretical advantages to this procedure are twofold: (1) The assessment of dimensionality is made using a model on which IRT is based and, (2) The measure of model misfit is directly related to the function minimized in the estimation procedure. From a practical perspective, the procedure is relatively inexpensive and fast (using ULS), and one does not encounter non positive-definite matrices common with the analysis

of tetrachoric correlation matrices (Hattie, 1984).

The main weakness in the technique is that there is no statistical significance test of the misfit of the model. In order to use the procedure, it is necessary to have some indication of approximate values on which to make decisions of fit or misfit. Studies should consider a variety of factors such as different test lengths, sample sizes, distribution of item parameters, etc,.

Therefore, the purposes of this study were to:

(1) Examine the distribution of the IFI in the null situation, (i.e., when the data are truly unidimensional);

(2) Examine the rejection rate of the IFI under various simulation conditions of a two-dimensional test structure;

(3) Compare the performance of the IFI with the T-statistic.

Methods

There were two parts to the study. The first part examined the distribution of the IFI_1 with randomly generated unidimensional data. The IFI_1 corresponding to the 95th percentile for each condition was determined and used as the critical value in the second part of the study. The purpose of the second part of this investigation was to determine the level of accuracy of the IFI_1 in detecting multidimensionality. In both sections of this study, data were randomly generated using

the general 2-parameter compensatory multidimensional model specified in equation 3.

$$P(i=1 \mid \underline{\theta}) = \frac{e^{1.7(\underline{a}'\underline{\theta} + d_i)}}{1 + e^{1.7(\underline{a}'\underline{\theta} + d_i)}} \qquad (3)$$

In the unidimensional case this model reduces to the usual logistic IRT model. Correlations between the latent abilities were set to be equal to zero in the generation of the two-dimensional data.

Unidimensional Data

In order to carry out the first part of this study, unidimensional data sets were generated with a modified version of M2PLGEN (Ackerman, 1987; modification by Gessaroli, 1990), a program designed to simulate binary response strings based on a two parameter logistic model. Two sample sizes were used (N=500 and N=1000). Discrimination parameters for the items were randomly generated from a Normal distribution with a mean and standard deviation of 1.0 and .25, respectively. By doing this, most of the item discrimination values fell between 0.4 and 1.6. Item difficulties were normally distributed ($N(0,1)$). The item difficulties were restricted to be between ±2.0. Test length was set to be either 15 or 45. Finally, data sets in each cell in this 2 x 2 design (sample size by test length) were replicated 100 times for a total of 400 unidimensional data sets. Each of

the 400 data sets was analyzed with both a 1-factor and 2-factor specification. In every case the IFI was calculated. The IFI corresponding to the 95th percentile in each cell was calculated and used as the critical value in the second part of the study.

Two-dimensional Data

In the second part of the study, two-dimensional data were generated and the unidimensionality was tested by calculating IFI_1 using the 1-factor and 2-factor SS_{Res}. These values were compared to the critical values determined in the first part of the study. Specifically, item difficulty and discrimination parameters similar to those considered by Berger & Knol (1990) were used to generate the multidimensional data sets. Again, as in the first part of the study, test lengths of 15 and 45 items were used as well as sample sizes of 500 and 1000. Two test structures were utilized reflecting different dimension strengths. The discrimination parameters used are shown in Table 1.

Insert Table 1 about here

The "weak" two-dimensional structure is designated by W2 whereas the "strong" two-dimensional structure is identified as S2. These patterns were repeated 5 times for the 15-item test length and 15 times for a test having 45 items. Item difficulty parameters of -2, -1, 0, 1, 2 were evenly distributed across the

different combinations of a_1 and a_2.

The IFI was calculated using NOHARMII (Fraser, 1983), a program based on McDonald's polynomial approximation to a normal ogive model. Because adequate starting values for the parameters to be estimated are essential for the minimization of the fit function, factor loadings obtained from a linear factor analysis of the matrix of phi-correlations among the items were used as these starting values. Stout's T statistic (Nandakumar, 1987) was computed using a program written by Junker (1988). Unidimensionality of the data sets was tested using the .05 level of significance.

Results

Unidimensional data sets

Descriptive statistics obtained for the IFI with various test lengths and sample sizes for the unidimensional data sets are presented in Table 2.

Insert Table 2 about here

As would be expected, the mean IFI for the 15 item data sets are larger than for the 45 item sets. There is more information available in the longer test resulting in a better estimation (i.e., smaller residual covariances) of the unidimensional

structure when estimating a 1-factor model. Thus, there is more information left to explain in the residual covariances with the shorter test length.

It appears as though the IFI_1 values corresponding to the 95th percentile do not differ appreciably with different sample sizes. However, the cutoffs are much smaller for the 45 item sets than for tests comprised of 15 items.

Table 3 displays the number of false rejections of unidimensionality using Stout's T statistic.

Insert Table 3 about here

It is clear from these results that the actual Type I error rate is close to the nominal α in all conditions simulated.

The results for both indicies, however, should be interpreted with caution given that they were based on only 100 replications and are specific to the simulation conditions used.

Multidimensional data sets

Table 4 shows the frequency of rejection rates of the assumption of unidimensionality for both the IFI and Stout's corrected T statistic when the data conformed to a "weak" two-dimensional structure.

Insert Table 4 about here

It appears that the IFI is fairly consistent in its ability to reject unidimensionality across sample sizes. Furthermore, its rejection rates are much more stable across test lengths compared to the T statistic, although, it does seem possible that longer tests will increase the rejection rate of the IFI.

The rejection rates of the T statistic, however, does seem to be very strongly influenced by both sample size and test length. Consistent with Nandakumar's (1987) results, the T statistic does not perform well with the 15 item test length. Its accuracy in rejecting unidimensionality does increase somewhat with the 45 item tests. However, the rejection rate in data sets having 1000 subjects is approximately twice that of data sets having sample sizes of 500. The rejection rates of the T statistic and the IFI are approximately equal in the 45 item tests having 1000 cases.

Table 5 presents results obtained with the "strong" two-dimensional data structures.

Insert Table 5 about here

It appears, from Table 5, that both the corrected T statistic and the IFI have a high degree of accuracy in rejecting the assumption of unidimensionality. Again, as with the weak two-dimensional structure, the T statistic is influenced by the length of the test. However, in this instance, sample size does

not seem to be an issue. The IFI rejected unidimensionality for every data set in all simulated conditions (available at the present moment) with the strong dimensionality structure.

Discussion

Results obtained for the T statistic parallel those from previous studies (Nandakumar, 1987; 1988; 1989; Stout, 1987). The accuracy of the statistic is greatly affected by test length and sample size. The utility of the T statistic increases as the length of the test increases. Again, these results support Nandakumar (1987) who states that the statistic should not be used when the test contains less than 25 items.

The IFI appeared to perform adequately in detecting multidimensionality of the test in all conditions simulated in the study.

There are several issues relating to the potential use of the IFI to assess test dimensionality. First, although the IFI did appear to perform quite well in this study, it is necessary to test the index under different conditions.

The IFI is based on the minimum of the fit function in the estimation procedure (ULS) used in NOHARMII. Although the sum of squares of the residual covariances is not the same as the mean absolute residual covariance used by Berger & Knol (1990), and is

the basis of Stout's essential dimensionality, it does, in principle, address the same issue. In keeping with the philosophy of essential dimensionality, however, the influence of a only a few multidimensional items on the IFI does need to be investigated. Berger and Knol (1990) indicate concerns that the mean SS_{Res} is sensitive to outliers. One would expect that this sensitivity, if it does exist, should decrease as the length of the test increases. However, an assessment of the "robustness" of the IFI to unimportant dimensions or items does seem necessary.

The IFI has the disadvantage of not having a statistical test of significance. Establishing a proper criterion by which to make a decision is somewhat arbitrary. Hopefully, the results of this study can, at least, provide some indication of the approximate magnitude of the IFI to be expected. Further examination of the IFI with different test lengths, sample sizes and dimension strengths would provide further insight into this problem. A scree plot of the IFI indices for subsequent dimensions, similar to those presented by Berger & Knol (1990) for the mean absolute residual covariances, might be useful. However, this approach also has the weakness of subjectivity in its interpretation. The scree plots were not examined in this study given that only estimates of one- and two-dimensional structures were examined and thus, only one IFI was calculated.

Summary

The IFI based on the SS_{Res} of the one and two-dimensional non-linear factor analyses of dichotomous test data did, in all cases, show fairly high rejection rates of unidimensionality when two-dimensional data were generated. Based on theses results it appears that this statistic has the potential to be used in the assessment of unidimensionality of test data and, more generally, in the determination of the number of dimensions underlying a test. Further studies investigating alternate test lengths, sample sizes and dimensionality structures, including those typifying essential dimensionality, must take place to provide a better understanding of the utility of the IFI.

Stout's T-statistic performed as expected. Its number of false rejections of unidimensionality was close to that predicted by the nominal significance level of the test. The T statistic seems best suited for long tests having large sample sizes. In these conditions, based on the results of this study, one would recommend the use of the T statistic. However, for smaller test lengths or smaller sample sizes, alternate indices such as the mean absolute residual covariance (Berger and Knol, 1990) or the IFI might be preferable.

References

Ackerman, T. A. (1985). M2PLGEN: A computer program for generating thetas and response strings corresponding to the M2PL model. Iowa City, Iowa: American College Testing.

Ackerman, T.A. (1987). A comparison study of the unidimensional IRT estimation of compensatory and noncompensatory multidimensional item response data. (Report No. 87-12). Iowa City, IA: The American College Testing Program.

Ansley, T.N., & Forsyth, R.A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. Applied Psychological Measurement, 9, 37-48.

Berger, M.P.F., & Knol, D.L. (1990, April). On the assessment of dimensionality in multidimensional Item Response Theory models. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA.

De Ayala, R.J., & Hertzog, M.A. (1989, March). A comparison of methods for assessing dimensionality for use in Item Response Theory. Paper presented at the annual meeting of the National Council on Measurement in Education, San

Francisco, CA.

Drasgow, P., & Parsons, C.K. (1983). Applications of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 7, 189-199.

Fraser, C. (1983). NOHARMII. A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. Armidale, Australia: The University of New England, Center for Behavioural Studies.

Goldstein, H., & Wood, R. (1989). Five decades of Item Response Modelling. British Journal of Mathematical and Statistical Psychology, 42, 139-167.

Green, S.B. (1983). Identifiability of spurious factors using linear factor analysis with binary items. Applied Psychological Measurement, 7(2), 139-147.

Hambleton, R.K., & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. Amherst, MA: University of Massachussets, Faculty of Education. (ERIC Document Reproduction Service No. ED 270 478)

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

McDonald, R.P. (1967). Nonlinear factor analysis. Psychometrika Monograph No. 15, 32(4, Pt. 2).

McDonald, R.P., & Ahlawat, K.S. (1974). Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 27, 82-99.

McDonald, R.P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

McDonald, R.P. (1989). Future directions for Item Response Theory. International Journal of Educational Research, 13(2), 205-220.

Nandakumar, R. (1987). Refinement of Stout's procedure for assessing latent trait dimensionality. Unpublished doctoral dissertation. Urbana-Champaign: University of Illinois.

Nandakumar, R. (1988, April). Modification of Stout's procedure for assessing latent trait unidimensionality. Paper

presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Nandakumar, R. (1989, March). Traditional dimensionality vs essential dimensionality. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Reckase, M.D., Carlson, J.E., Ackerman, T.A., & Spray, J.A. (1986, June). The interpretation of unidimensional IRT parameters when estimated from multidimensional data. Paper presented at the annual meeting of the Psychometric Society, Toronto, Ont.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. Psychometrika, 52(4), 589-617.

Takane, Y., & De Leeuw, J. (1987). On the relationship between Item Response Theory and Factor Analysis of discretized variables. Psychometrika, 52, 393-408.

Zwick, R.W., & Velicer, W.F. (1986). Comparison of five rules for determining the number of components to retain. Psychological Bulletin, 99(3), 432-442.

Table 1

Item Discrimination Parameters Defining the Two-Dimensional Structures

W2		S2	
a_1	a_2	a_1	a_2
1.0	0.0	2.0	2.0
1.0	0.5	0.0	2.0
0.0	0.5	2.0	0.0

Table 2

Descriptive Statistics For the IFI : Unidimensional Data Sets

Test Length	15 Items		45 Items	
Sample Size	500	1000	500	1000
Mean	.253	.273	.114	.099
SD	.076	.075	.021	.027
Skewness	.154	.629	.756	.863
Kurtosis	.100	.664	1.352	1.154
PR_{95}	.363	.400	.159	.149

Table 3

Number of Rejections of Unidimensionality Using the T Statistic Per 100 Trials: Unidimensional Data sets

Test Length	Sample Size	T Statistic
15	500	2
	1000	3
45	500	2
	1000	4

Table 4

Number of Rejections of unidimensionality per 100 Trials:

"Weak" Two-Dimensional Data Sets

Test Length	Sample Size	IFI	T Statistic
15	500	65	4
	1000	67	4
45	500	74	38
	1000	79	77

Table 5

Number of Rejections of Unidimensionality per 100 Trials: "Strong" Two-Dimensional Data Sets

Test Length	Sample Size	IFI	T Statistic
15	500	100	71
	1000	100	77
45	500	N/A	100
	1000	N/A	100